Vidyavardhaka Sangha®, Mysore
VIDYAVARDHAKA COLLEGE OF ENGINEERING
Autonomous Institute, affiliated to Visvesvaraya Technological University, Belagavi
(Approved by AICTE, New Delhi & Government of Karnataka)
Accredited by NBA (CV, CS, EE, EC, IS & ME) | NAAC with 'A' Grade
P.B. No. 206, Gokulam III Stage, Mysuru-570 002, Karnataka, India
Phone: +91 821 4276201 /202 /225, Fax: +91 824 2510677
Web: http://www.vvce.ac.in
@vvceofficial

# *Course Content & Planning*

| SEMESTER – II | | | |
|---|---|---|---|
| **Course Name** | **: Data Science** | **Course Code :** | **MDSCCS202** |
| **Number of Lecture Hours / Week** | **: 03** | **CIE Marks :** | **50** |
| **Number of Tutorial / Practical Hours / Week** | **: 02** | **SEE Marks :** | **50** |
| **Total Number of Lecture + Tutorial/Practical Hours** | **: 40+24=64** | **SEE Duration :** | **03 Hours** |
| **L:T:P** | **: 3:0:2** | **CREDITS :** | **04** |

**COURSE PREREQUISITES:**
Basic knowledge of working with python commands, logical thinking and problem solving skills are required to learn the course.

**COURSE OVERVIEW**
Data is one of the important features of every organization because it helps business leaders to make decisions based on facts, statistical numbers and trends. Usage of scientific approaches, procedure, algorithms, and framework to extract the knowledge and insight from a huge amount of data is much needed in today's world. This course on Data Science introduces to the application of scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data and apply knowledge and actionable insights from data across a broad range of application domains.

**COURSE LEARNING OBJECTIVES (CLO)**
- Get familiarized with NumPy, Pandas and Matplotlib packages available in python
- Demonstrate the data manipulation techniques using Pandas
- Demonstrate the data visualization techniques using Matplotlib and Seaborn
- Explore ML algorithms for data analysis

| MODULES | TEACHING HOURS |
|---|---|
| **MODULE 1**<br>**Introduction to NumPy:** Understanding Data Types in Python, The Basics of NumPy Arrays, Computation on NumPy Arrays: Universal Functions, Aggregations: Min, Max, Computation on Arrays: Broadcasting, Comparisons, Masks, and Boolean Logic, Fancy Indexing, Sorting Arrays, Structured Data: NumPy's Structured Arrays. | **8** |
| **MODULE 2**<br>**Data Manipulation with Pandas:** Introducing Pandas Objects, Data Indexing and Selection, operating on Data in Pandas, Handling Missing Data, Hierarchical Indexing, Combining Datasets, Aggregation and Grouping, Pivot Tables, Working with Time Series. | **8** |
| **MODULE 3**<br>**Visualization with Matplotlib and Seaborn:** Simple Line Plots, Simple Scatter Plots, Visualizing Errors, Density and Contour Plots, Histograms, Binnings, and Density plots, Customizing plot legends, Multiple sub plots, visualizing with Seaborn, Seaborn vs Matplotlib, Exploring Seaborn plots. | **8** |

| | |
|---|---|
| **MODULE 4**<br>**ML using Scikit Learn-1:** Categories of machine learning and their qualitative examples, Introduction to Scikit-Learn, Hyper parameters and model validation, Bayesian classification, Gaussian Naïve Bayes, Multinomial Naïve Bayes classification. | **8** |
| **MODULE 5**<br>**ML using Scikit Learn-2:** Introduction to Decision Trees and random forests, Appropriate problems for decision tree learning, Basic decision tree algorithm, Issues in decision tree learning, ensembles and estimators in random forests, Introduction to K- Means clustering, Expectation and maximization in clustering, weaknesses of k-means clustering | **8** |
| **PRACTICAL MODULE**<br><br>1. Read the given data "churn.csv" and save it as a dataframe called churn_data. Perform following operations on the dataframe i) Count total number of duplicate records in the dataframe ii) Count the no. of duplicate records in the churn dataframe based on the cutomerID column iii) Count number of missing values in each columns iv) Count the total no. of missing values for the variable TotalCharges v) Average monthly charge paid by a customer for the services he/she has signed up for vi) Display the records having "1@#" under the variable Dependents. Vii) Replace null values in churn dataframe by median value or by max count class category. *(https://drive.google.com/file/d/1JSYGIIkIZr4jyheDEH0X1_TMSnQ2CzXc/view)*<br>2. Using the data on births in the United States, provided by the Centers for Disease Control (CDC), Find i) Total number of US births by year and gender ii) Average daily births by day of week and decade iii) Average daily births by date. *(https://raw.githubusercontent.com/jakevdp/data-CDCbirths/ master/births.csv)*<br>3. Explore the bicycle counts on Seattle's Fremont Bridge Data with respect to i) Hourly bicycle counts on Seattle's Fremont bridge ii) Weekly bicycle crossings of Seattle's Fremont bridge iii) Average daily bicycle counts iv) Average hourly bicycle counts by weekday and weekend. *(https://data.seattle.gov/api/views/65db-xm6k/rows.csv?accessType=DOWNLOAD).*<br>4. Visualize and understand finishing results from a marathon race with respect to distribution of split fractions, distribution of split fractions by gender and age, split fraction vs finishing time by gender. *(https://raw.githubusercontent.com/jakevdp/marathon-data/master/marathon-data.csv)*<br>5. Using the sparse word count features from the Newsgroups corpus data set, classify short documents to different categories.<br>6. Identify similar handwritten without using the original label information using clustering technique.<br>7. Locate and identify characters in images of handwritten digits using a random forest classifier.<br><br>**Open Ended Experiments**<br><br>1. For any dataset from UCI Machine repository, explore the relationship between the features using Pandas framework and Seaborn visualization library. Summarize the inferences from the data. | **24** |

2. Design and develop a classifier to recognize human faces. Test the classifier and plot the confusion matrix for the same.

Select any dataset from UCI Machine repository which is suitable for classification task. Build a suitable classifier for the same. Test the model and summarize the inferences from the results obtained.

## Textbooks

1. Jake VanderPlas: Python Data Science Handbook. O'Reilly publication 2017.
2. Tom M. Mitchell, "Machine Learning", McGraw Hill publication, 1997.

## Reference Books

1. Wes Mckenny: Python for Data Analysis, O'Reilly publication 2013.
2. Charles R. Severance, "Python for Everybody: Exploring Data Using Python 3", 1st Edition, CreateSpace Independent Publishing Platform, 2016.

## COURSE OUTCOMES (COs)

At the end of the course students will be able to

| | |
|---|---|
| ***CO1*** | Explain data preprocessing, data visualization, prediction and clustering tasks |
| ***CO2*** | Apply different libraries for data manipulation, data visualization and machine learning tasks |
| ***CO3*** | Analyze a dataset using exploratory data analysis techniques |
| ***CO4*** | Analyze and present a substantial technical content |

**Course Assessment Plan**

| CO | Marks Distribution | | | | | Total Marks | Weightage (%) |
|---|---|---|---|---|---|---|---|
| | Module- 1 | Module- 2 | Module- 3 | Module- 4 | Module- 5 | | |
| CO1 | 12 | | 10 | 8 | | 30 | 30 |
| CO2 | 8 | 12 | 10 | 7 | 12 | 49 | 49 |
| CO3 | | 8 | | 5 | 8 | 21 | 21 |
| | **20** | **20** | **20** | **20** | **20** | **100** | |

| Intra-Module Choice for Answering of Questions | |
|---|---|
| **Module-II** | **Module-IV** |
| In other modules there will be no choice for answering | |